

Rectified Gaussian Scale Mixtures and the Sparse Non-Negative Least Squares Problem

Alican Nalci, Igor Fedorov, and Bhaskar D. Rao

Abstract—In this paper we introduce a hierarchical Bayesian framework to obtain sparse and non-negative solutions to the sparse non-negative least squares problem (S-NNLS). We introduce a new family of scale mixtures, the Rectified Gaussian Scale Mixture (R-GSM), to model the sparsity enforcing prior distribution for the signal of interest. One advantage of the R-GSM prior is that through proper choice of the mixing density it encompasses a wide variety of heavy tailed distributions, such as the rectified Laplacian and rectified Student's t distributions. Similar to the Gaussian Scale Mixture (GSM) approach, a Type II Expectation-Maximization framework is developed to estimate the hyper-parameters and obtain a point estimate of the parameter of interest. In the proposed method, called rectified Sparse Bayesian Learning (R-SBL), we provide two ways to perform the Expectation step; Markov-Chain Monte-Carlo (MCMC) simulations and a simple yet effective diagonal approximation approach (DA). Through numerical experiments we show that R-SBL outperforms existing S-NNLS solvers in terms of both signal and support recovery and that the proposed DA approach admits both computational efficiency and numerical accuracy.

Index Terms—Non-negative Least Squares, Bayesian Sparse Signal Recovery, Rectified Gaussian Scale Mixtures, Diagonal Approximation, Markov-Chain Monte-Carlo

I. INTRODUCTION

In this paper we consider a linear system of equations of the form

$$\mathbf{y} = \Phi \mathbf{x} + \mathbf{v} \quad (1)$$

where the solution of interest, $\mathbf{x} \in \mathbb{R}_+^M$, is assumed to be non-negative. The matrix $\Phi \in \mathbb{R}^{N \times M}$ is fixed and related to the underlying physical problem, $\mathbf{y} \in \mathbb{R}^N$ is the measurement vector, and \mathbf{v} is the additive measurement noise modeled as a zero mean Gaussian random vector with uncorrelated entries, i.e. $\mathbf{v} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. Recovering the optimal solution to (1) is known as solving the non-negative least squares (NNLS) problem. NNLS has received considerable attention in the context of methods for solving systems of linear equations [1], density estimation [2], compressive non-negative imaging [3], and non-negative matrix factorization (NMF) [4], [5], [6], among others. Non-negative data is also a natural occurrence in many application areas, such as text mining [7], image

processing [8], speech enhancement [9], and spectral decomposition [10], [11] and requires special consideration. To this end, we pose the solution to the NNLS problem in (1) as

$$\underset{\mathbf{x} \geq 0}{\text{minimize}} \quad \|\mathbf{y} - \Phi \mathbf{x}\|_2. \quad (2)$$

In many applications, $N < M$ and, therefore, (1) is under-determined. As a result, a unique solution to (2) may not exist. However, recovering an exact solution is possible if additional information is known about the solution. A useful assumption, and one that has been recently made in many applications, is that the solution is sparse [12], [13], [14]. The problem becomes more well-posed if \mathbf{x} is known to be sparse, i.e. has few non-zero elements. In this case, (2) is simply modified to

$$\underset{\mathbf{x} \geq 0, \mathbf{y} = \Phi \mathbf{x}}{\text{minimize}} \quad \|\mathbf{x}\|_0 \quad (3)$$

where $\|\mathbf{x}\|_0$, the ℓ_0 pseudo-norm, is the number of non-zero elements in \mathbf{x} . We will refer to (3) as the sparse NNLS (S-NNLS) problem.

Directly solving (3) is not tractable, since the ℓ_0 pseudo-norm is not convex in its arguments and the problem is NP-hard in general [15]. Therefore greedy algorithms have been suggested to approximate the solution [15], [16]. One example is the class of algorithms known as Orthogonal Matching Pursuit (OMP) [17], which successively select non-zero elements of \mathbf{x} in a greedy fashion. In order to adapt OMP to the S-NNLS problem, the criterion by which a new non-zero element of \mathbf{x} is selected is modified to select the one having the largest *positive* value [18]. Another approach in this class of algorithms first finds an \mathbf{x} such that $\|\mathbf{y} - \Phi \mathbf{x}\|_2 \leq \epsilon$ and $\mathbf{x} \geq 0$ using the active-set Lawson-Hanson algorithm [1] and then prunes \mathbf{x} with a greedy procedure until $\|\mathbf{x}\|_0 \leq K$, where K is a pre-specified desired sparsity [4].

Greedy algorithms are computationally attractive to solve NP-hard problems, but often lead to sub-optimal solutions [15]. Thus, numerous convex relaxations of the ℓ_0 penalty have been considered as an alternative to greedy methods [15]. One simple alternative is to replace the ℓ_0 penalty with ℓ_1 and cast the problem in (3) as

$$\underset{\mathbf{x} \geq 0}{\text{minimize}} \quad \|\mathbf{y} - \Phi \mathbf{x}\|_2 + \lambda \|\mathbf{x}\|_1 \quad (4)$$

where $\lambda > 0$ is a suitably chosen regularization parameter. The advantage of this formulation is that it is in the form of a constrained convex optimization problem and can be solved by any number of methods [19], [20]. One approach is to estimate \mathbf{x} through a projected gradient descent procedure [21]. Whereas the projected gradient descent procedure requires a line-search in order to calculate the learning rate at

The authors are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92092 USA (analci@eng.ucsd.edu; ifedorov@eng.ucsd.edu; brao@ucsd.edu). This work was partially supported by KIET of MOTIE grant funded by Korea government in 2015. [No. 10050527, General-purpose module, sensor and system development projects to enhance industrial safety network]. Alican Nalci was partially supported by the UCSD Frontiers of Innovation Scholars Program. Igor Fedorov was partially supported by the ARCS Foundation.

each iteration, the approach can be simplified by employing a multiplicative update rule which guarantees a decrease in (4) at each iteration [22]. In fact, the ℓ_1 norm penalty in (4) can be replaced by any sparsity inducing regularization function $g(\mathbf{x})$:

$$\underset{\mathbf{x} \geq 0}{\text{minimize}} \quad \|\mathbf{y} - \Phi \mathbf{x}\|_2 + \lambda g(\mathbf{x}). \quad (5)$$

For instance, [23], [24] consider $g(\mathbf{x}) = \sum_{i=1}^M \log(x_i^2 + \beta)$, which leads to an iterative re-weighted optimization approach.

An alternative and more promising view on the S-NNLS problem is to cast the entire problem in a Bayesian framework and consider the maximum a-posteriori (MAP) estimate of \mathbf{x} given the data \mathbf{y} :

$$\mathbf{x}_{MAP} = \arg \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}). \quad (6)$$

There is a strong connection between the Bayesian framework in (6) and deterministic formulations like the one in (5): it has recently been shown that formulations of the form in (5) are equivalent to the formulation in (6) with the proper choice of prior $p(\mathbf{x})$ [25]. For example, considering a separable $p(\mathbf{x})$ of the form

$$p(\mathbf{x}) = \prod_{i=1}^M p(x_i), \quad (7)$$

the ℓ_1 regularization approach in (4) is equivalent to the Bayesian approach in (6) with an exponential prior for $p(x_i)$. In this paper, our emphasis will be on Bayesian approaches to solving (1).

A. Contributions of the Paper

- We introduce a novel class of sparsity inducing priors for the S-NNLS problem, namely priors modeled as a rectified Gaussian Scale Mixture (R-GSM). These are an extension of the Gaussian Scale Mixture (GSM) model [26], [27], [28], [29] and represent a large class of heavy tailed priors appropriate for inducing sparsity.
- We cast the S-NNLS problem in a Bayesian framework using R-GSM priors and develop a Type II inference procedure, which we call Rectified Sparse Bayesian Learning (R-SBL). We develop a detailed Bayesian inference procedure for estimating \mathbf{x} using the Expectation-Maximization (EM) algorithm using both MCMC and a Diagonal Approximation (DA) approach that performs very close to MCMC in terms of recovery performance but is computationally much more efficient.
- We detail two potential point estimate strategies for R-SBL to find the optimal solution: the mean and mode point estimates.
- We demonstrate the utility of the R-GSM prior and efficacy of the R-SBL algorithm with extensive experimental results comparing the proposed technique to existing S-NNLS algorithms. We also discuss the robustness of the proposed method in the case of noisy measurements.

B. Organization of the Paper

In the following sections, we provide details of the proposed R-SBL algorithm. In Section II, we discuss the advantages of scale mixture priors for $p(\mathbf{x})$ and introduce the R-GSM prior. In Section III, we define the Type I and Type II Bayesian approaches to solve the S-NNLS problem and introduce the main R-SBL framework with R-GSM prior. We provide details of an EM technique for estimating \mathbf{x} using MCMC and our DA method in Section III-B. We present experimental results comparing the proposed R-SBL algorithm to existing methods in Section IV. Finally, we conclude the work in Section V.

II. RECTIFIED GAUSSIAN SCALE MIXTURES

In this work, we assume separable priors of the form (7) and focus on the choice of $p(x_i)$. The choice of prior is very important because it plays a central role in the Bayesian inference procedure. For the problem at hand, the prior must be sparsity inducing, satisfy the non-negativity constraint, be general and versatile enough to be widely applicable. Consequently, we consider the scale mixture prior:

$$p(x_i) = \int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i. \quad (8)$$

Scale mixture priors were first considered in the context of GSM's [26]. It is well known that super-gaussian densities are the best priors for promoting sparsity [30], [31] and most such priors can be represented in the form shown in (8), with the proper choice of $p(\gamma_i)$ [32], [33], [29], [28], [27]. This has made GSM based priors a valuable form of prior for the general sparse signal recovery problem. Another advantage of the scale mixture prior is that it establishes a Markovian structure of the form

$$\gamma \rightarrow \mathbf{x} \rightarrow \mathbf{y}$$

where inference can be performed in the \mathbf{x} domain (MAP or Type I) and also in the γ domain (Type II). This provides additional flexibility and opportunity to explore algorithm development. Experimental results in the standard sparse signal recovery problem show that performing inference in the γ domain consistently achieves superior performance [25], [30]. The performance gains may be due to the intuition that γ is deeper in the Markovian chain than \mathbf{x} , so the influence of errors in performing inference in the γ domain may be diminished [25], [34], [35]. At the same time, γ is close enough to \mathbf{y} in the Markovian chain such that meaningful inference about γ can still be performed [25].

Although priors of the form shown in (8) have been used widely in the compressed sensing literature (where the signal model is identical to (1) without the non-negativity constraint) [25], this framework has not been extended to the S-NNLS problem. In [36], a Rectified Gaussian (RG) prior is considered for $p(x_i)$ within a variational bayesian inference framework. Our goal in this work is to develop a more flexible and general class of priors. Considering the findings that the scale mixture prior is superior to most existing sparse signal recovery algorithms [25], we propose a R-GSM prior for the S-NNLS problem, where $p(x_i|\gamma_i)$ in (8) is given by the RG

distribution. We refer to the proposed inference framework with R-GSM prior as Rectified Sparse Bayesian Learning (R-SBL).

The RG distribution is defined as

$$\mathcal{N}^R(x|\mu, \gamma) = \sqrt{\frac{2}{\pi\gamma}} \frac{1}{\operatorname{erfc}\left(-\frac{\mu}{\sqrt{2\gamma}}\right)} e^{-\frac{(x-\mu)^2}{2\gamma}} u(x)$$

where μ is the location parameter, γ is the scale parameter, $u(x)$ is the unit step function, and $\operatorname{erfc}(x)$ is the complementary error function, defined as

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt.$$

When $\mu = 0$, the RG density becomes

$$\mathcal{N}^R(x|0, \gamma) = 2\mathcal{N}(x|0, \gamma)u(x) = \sqrt{\frac{2}{\pi\gamma}} e^{-\frac{x^2}{2\gamma}} u(x). \quad (9)$$

As noted in [37], [38], closed form inference computations using scale mixtures of RG's is tractable only if the location parameter is zero, which effectively makes the $\operatorname{erfc}(\cdot)$ vanish. Thus, in this work, we focus on R-GSM with $\mu = 0$.

Motivated by Gaussian Scale mixtures, we use the RG prior in a scale mixture framework in order to provide a general and flexible class of priors and to better promote *sparse* non-negative solutions. R-GSM priors have the form

$$p(x) = \int_0^\infty \mathcal{N}^R(x|0, \gamma) p(\gamma) d\gamma. \quad (10)$$

Different choices of $p(\gamma)$ lead to different choices of priors. A few example are presented below.

A. Examples of R-GSM Representation of Sparse Prior

A random variable modeled by a R-GSM can also be viewed as generating a random variable with a GSM and then taking its absolute value. This allows one to leverage the GSM prior literature to develop the class of R-GSM priors.

For instance, consider the rectified Laplacian, prior for x :

$$p_e(x) = \lambda e^{-\lambda x} u(x).$$

By using an exponential prior for $p(\gamma)$,

$$p(\gamma) = \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \gamma}{2}} u(\gamma) \quad (11)$$

we can express $p_e(x)$ in the R-GSM form [39] as

$$\begin{aligned} p_e(x) &= 2u(x) \int_0^\infty \mathcal{N}(x|0, \gamma) \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \gamma}{2}} u(\gamma) d\gamma \\ &= \lambda e^{-\lambda x} u(x). \end{aligned}$$

Likewise, by considering the R-GSM with $p(\gamma)$ given by the Gamma(a, b) distribution, we get a rectified student-t distribution for $p(x)$. In this case, (8) simplifies to [30]

$$\begin{aligned} p(x) &= 2u(x) \int_0^\infty \mathcal{N}(x|0, \gamma) \frac{\gamma^{a-1} e^{-\frac{\gamma}{b}}}{a^b \Gamma(a)} d\gamma \\ &= \frac{2b^a \Gamma(a + \frac{1}{2})}{(2\pi)^{\frac{1}{2}} \Gamma(a)} \left(b + \frac{x^2}{2}\right)^{-(a + \frac{1}{2})} u(x) \end{aligned}$$

where Γ is defined as

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt.$$

More generally, all of the distributions represented by the GSM have a corresponding rectified version represented by a R-GSM, e.g. contaminated Normal and slash, symmetric stable and logistic, hyperbolic, etc. [29], [28], [27], [26], [32] [33].

III. BAYESIAN INFERENCE WITH SCALE MIXTURE PRIOR

There are two Bayesian approaches for solving the S-NNLS problem with a scale mixture prior. The first approach, called Type I estimation (MAP) performs inference in the \mathbf{x} domain by treating γ as the hidden variable. On the other hand, the second approach, called Type II, performs inference in the γ domain. We briefly discuss Type I in the next section and the major portion of the paper is dedicated to Type II estimation because of its observed superiority in standard sparse signal recovery problems [25].

A. Type I or MAP estimation

Employing Type I to solve S-NNLS translates into calculating the maximum a-posteriori (MAP) estimate of \mathbf{x} given \mathbf{y} :

$$\arg \min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{x}\|_2^2 - \lambda \sum_{i=1}^M \log p(x_i). \quad (12)$$

Many of the ℓ_0 relaxation methods described in Section I can be re-derived from the Type I perspective. For instance, by choosing an exponential prior for $p(\gamma_i)$, leading to an exponential prior on $p(x)$, (12) reduces to the ℓ_1 regularization approach in (4) with the added benefit that λ has a probabilistic interpretation. Similarly, by choosing a Gamma prior for $p(\gamma_i)$, which corresponds to a rectified student-t prior on $p(x_i)$, (12) reduces to

$$\arg \min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{x}\|_2^2 + \lambda \sum_{i=1}^M \log \left(b + \frac{x_i^2}{2}\right)$$

which leads to the re-weighted ℓ_2 approach to the S-NNLS problem described in [23][24].

B. Type II Estimation: R-SBL for S-NNLS

The R-SBL framework aims to infer the hyper-parameters γ that shape the R-GSM prior. A MAP estimate of γ is computed and then the posterior density of interest $p(\mathbf{x}|\mathbf{y})$ is approximated as $p(\mathbf{x}|\mathbf{y}, \gamma_{MAP})$. Having estimated the posterior density, appropriate point estimates can be readily obtained.

Based on past experience in Type II estimation for the standard sparse signal recovery problem, several strategies exist for estimating γ . The first strategy considers the problem of forming a MAP estimate of γ given \mathbf{y} by directly minimizing the appropriate posterior density. In this case, $p(\gamma|\mathbf{y})$ does not appear to admit a closed form, so we do not pursue this strategy. The second strategy, investigated here, aims to estimate γ by utilizing an EM framework.

In the EM approach, we treat $(\mathbf{x}, \mathbf{y}, \gamma)$ as the complete data and \mathbf{x} as the hidden variable. The E-step involves determining $Q(\gamma, \gamma^t)$, which is defined as the conditional expectation of the complete data log-likelihood:

$$Q(\gamma, \gamma^t) = E_{\mathbf{x}|\mathbf{y}; \gamma^t, \sigma^2} [\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}|\gamma) + \log p(\gamma)] \quad (13)$$

$$\doteq \sum_{i=1}^M E_{\mathbf{x}|\mathbf{y}; \gamma^t, \sigma^2} \left[-\frac{1}{2} \log \gamma_i - \frac{x_i^2}{2\gamma_i} \right] \quad (14)$$

where t refers to the iteration index and \doteq refers to the fact that constant terms and terms which don't depend on γ have been omitted since they don't effect the maximization step. For the sake of simplicity of this paper, we also assume a non-informative prior on γ [30]. In the M-step, we maximize $Q(\gamma, \gamma^t)$ with respect to γ by taking the derivative and setting it equal to zero, which yields an estimate of γ :

$$\gamma_i^{t+1} = E_{\mathbf{x}|\mathbf{y}; \gamma^t, \sigma^2} [x_i^2] := \langle x_i^2 \rangle \quad (15)$$

To compute $\langle x_i^2 \rangle$, we first consider the form of the posterior $p(\mathbf{x}|\mathbf{y}, \gamma, \sigma^2)$:

$$p(\mathbf{x}|\mathbf{y}, \gamma) = c(\mathbf{y}) e^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}} u(\mathbf{x}) \quad (16)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by [40], [30], [41] as

$$\boldsymbol{\mu} = \boldsymbol{\Gamma} \boldsymbol{\Phi}^T (\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \boldsymbol{\Gamma} \boldsymbol{\Phi}^T)^{-1} \mathbf{y} \quad (17)$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma} \boldsymbol{\Phi}^T (\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \boldsymbol{\Gamma} \boldsymbol{\Phi}^T)^{-1} \boldsymbol{\Phi} \boldsymbol{\Gamma} \quad (18)$$

where $\boldsymbol{\Gamma} = \text{diag}(\gamma)$. The posterior in (16) is known as a multivariate RG (or multivariate truncated normal [42]).

The normalizing constant $c(\mathbf{y})$ does not appear to have closed form. The multivariate RG is a particularly difficult distribution to work with because its marginals are not univariate-RG's and do not admit a closed form [42], which makes computing the marginal moments difficult. As a result, we resort to calculating $\langle x_i^2 \rangle$ with two strategies. The first strategy uses MCMC to draw samples from $p(\mathbf{x}|\mathbf{y}, \gamma)$ and estimate $\langle x_i^2 \rangle$ from those samples. This process can be time intensive but is guaranteed to generate quality estimates. The second strategy is an approximation to the true second moments of the posterior $p(\mathbf{x}|\mathbf{y}, \gamma, \sigma^2)$. In this approach, in the moment calculations for $\langle x_i^2 \rangle$ we assume that the off-diagonal terms of $\boldsymbol{\Sigma}$ are very close to zero hence we disregard them. This leads to a very fast algorithm that uses closed form computations and outperforms baseline S-NNLS solvers by a large margin. We call this the diagonal approximation or (DA) for short.

1) *Markov-Chain Monte-Carlo*: In this work, we employ a Gibbs sampling [43] strategy, which is a popular MCMC technique for sampling from multivariate distributions. We follow [44] and begin by introducing the transformation

$$\mathbf{w} = \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \quad (19)$$

where \mathbf{L} is the lower triangular cholesky decomposition of $\boldsymbol{\Sigma}$. It can be shown that \mathbf{w} has a truncated normal distribution $\text{TN}(\mathbf{w}; 0, \mathbf{I}, \mathbf{L}, -\boldsymbol{\mu}, \infty)$ as in [44], where

$$\text{TN}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{\mathbf{R}}, \boldsymbol{\alpha}_L, \boldsymbol{\alpha}_U) = \quad (20)$$

$$\left(\frac{c_{tn} e^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}}}{c_{tn} e^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2}}} \right) \mathbf{1}_{\boldsymbol{\alpha}_L \leq \hat{\mathbf{R}} \mathbf{w} \leq \boldsymbol{\alpha}_U} \quad (21)$$

and $\mathbf{1}(\cdot)$ is the indicator function and c_{tn} is the normalizing constant.

The Gibbs sampler then proceeds by iteratively drawing samples from the conditional distribution $p(w_i | \mathbf{y}, \gamma, \sigma^2, \mathbf{w}_{-i})$, where \mathbf{w}_{-i} refers to the vector containing all but the i 'th element of \mathbf{w} . Given a set of samples drawn from \mathbf{w} , we can generate samples from the original distribution of interest by inverting the transformation:

$$\{\mathbf{x}^n\}_{n=1}^N = \{\mathbf{L} \mathbf{w}^n + \boldsymbol{\mu}\}_{n=1}^N \quad (22)$$

Finally, sample $\langle \widetilde{x_i^2} \rangle$ can be calculated using

$$\langle \widetilde{x_i^2} \rangle = \frac{1}{N} \sum_{n=1}^N (x_i^n)^2 \quad (23)$$

2) *Diagonal Approximation (DA)*: As an approximation to the true moments, we propose that the moments $\langle x_i \rangle$ and $\langle x_i^2 \rangle$ will be largely affected by the diagonal entries of $\boldsymbol{\Sigma}$. Therefore, inspired by the diagonal approximation ideas in [45], [46], we set $\Sigma_{ij} = 0$ for $i \neq j$ just for moment calculations. Through numerical experiments we verify that this assumption is a very good one and produces very good estimates for $\langle x_i \rangle$ and $\langle x_i^2 \rangle$ outperforming S-NNLS solvers in terms of sparse recovery performance and performing fairly similar to MCMC.

More importantly, MCMC procedure is computationally very intensive especially for large problem sizes and convergence of $\langle x_i^2 \rangle$ appears to be slow. DA approach on the other hand requires a single closed-form update equation requiring only the computation of the complimentary error function.

Carrying out the second moment computation with this assumption leads to the following approximation of $\langle x_i^2 \rangle$:

$$\langle x_i^2 \rangle \approx \mu_i^2 + \Sigma_{ii} + \mu_i \sqrt{\frac{1}{\pi}} \sqrt{\Sigma_{ii}} \frac{e^{-\frac{\mu_i^2}{2\Sigma_{ii}}}}{\text{erfc}\left(-\frac{\mu_i}{\sqrt{2\Sigma_{ii}}}\right)} \quad (24)$$

where $\text{erfc}(\cdot)$ refers to the complimentary error function. This is simply the second moment of a univariate RG density given in [47].

3) *Point estimate of \mathbf{x}* : The above equations are sufficient to implement the exact MCMC and DA versions of the EM algorithm. After finding an estimate of γ , $\hat{\gamma}$, the goal is to find a point estimate of \mathbf{x} from $p(\mathbf{x}|\mathbf{y}, \hat{\gamma}, \sigma^2)$. The optimal estimator of \mathbf{x} in the mean-square-error (MSE) sense is simply given by [48]

$$\hat{\mathbf{x}}_{\text{mean}} = E_{\mathbf{x}|\mathbf{y}, \hat{\gamma}}[\mathbf{x}] \quad (25)$$

where the expectation can be computed by using MCMC to sample from $p(\mathbf{x}|\mathbf{y}, \hat{\gamma}, \sigma^2)$ and average the samples or

Require: $\mathbf{y}, \Phi, \sigma^2, \epsilon$

- 1: Initialize $\Gamma^0 = \mathbf{I}$, $t = 1$
- 2: **while** $\frac{\|\gamma^{t-1} - \gamma^t\|_2}{\|\gamma^{t-1}\|_2} \geq \epsilon$ **do**
- 3: Calculate μ^t using (17) and Γ^{t-1}
- 4: Compute Σ^t using (18) and Γ^{t-1}
- 5: Option 1: Calculate γ^{t+1} using MCMC in (23)
- 6: Option 2: Approximate γ^{t+1} using DA in (24)
- 7: $t \leftarrow t + 1$
- 8: **end while**
- 9: For the mean estimator, compute $E_{\mathbf{x}|\mathbf{y}, \gamma^t, \sigma^2}[\mathbf{x}]$ using the MCMC sample mean or using the DA in (26)
- 10: For the mode estimator compute $\hat{\mathbf{x}}_{mode}$ using (27) with the converged values of $\gamma^{t_{end}}$.

Fig. 1: R-SBL Algorithm

by using the diagonal approximation, which results in the approximation i.e. the first moment for univariate RG in [47].

$$E_{\mathbf{x}|\mathbf{y}, \hat{\gamma}}[x_i] \approx \mu_i + \sqrt{\frac{2}{\pi}} \sqrt{\Sigma_{ii}} \frac{e^{-\frac{\mu_i^2}{2\Sigma_{ii}}}}{\text{erfc}\left(-\frac{\mu_i}{\sqrt{2\Sigma_{ii}}}\right)} \quad (26)$$

An alternative point estimate is to use $\hat{\mathbf{x}}_{mode}$, given by

$$\begin{aligned} \hat{\mathbf{x}}_{mode} &= \arg \max_{\mathbf{x}} p(\mathbf{x} | \mathbf{y}, \hat{\gamma}, \sigma^2) \\ &= \arg \min_{\mathbf{x} \geq 0} \|\mathbf{y} - \Phi \mathbf{x}\|_2^2 + \lambda \sum_{i=1}^M \frac{x_i^2}{\gamma_i} \end{aligned} \quad (27)$$

where (27) can be solved by any NNLS solver. $\hat{\mathbf{x}}_{mode}$ is a favorable point estimate because it chooses the peak of $p(\mathbf{x} | \mathbf{y}, \hat{\gamma}, \sigma^2)$, which could be multi-modal and not characterized well by its mean. Note that if all entries of $\hat{\mu}$ are positive, then calculating (27) is not necessary as this point is the true mode, i.e. $\hat{\mathbf{x}}_{mode} = \hat{\mu}$. The performance of both point estimate methods is discussed in Section IV.

The complete R-SBL algorithm is summarized in Figure 1 for easy reference.

IV. EXPERIMENTS

In this section we provide two sets of experiments to show the recovery performance of the R-SBL approach.

The first set presents the MCMC results along with the DA and the baseline S-NNLS solvers. This experiment aims to compare MCMC with DA and establishes the superiority of R-SBL over the baseline methods. We focus on two relatively smaller sized problems (P1) and (P2) as MCMC is computationally not feasible for large problems.

The second set contains extensive simulations comparing the proposed DA approach with baseline S-NNLS solvers for a larger problem size and for various cardinalities. In both first and second sets, we simulate a near 'noiseless' case where the noise has distribution $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma^2 = 10^{-8}$.

We present results of the proposed DA for a wide variety of distributions for \mathbf{x} and Φ . Additionally, in a separate simulation, we set the noise variance such that the signal to

noise ratio (SNR) is 20 dB to test the robustness of the R-SBL with DA under noisy conditions.

In all of the experiments, we compare our approach to available baseline S-NNLS solvers: SLEP- ℓ_1 [49], and NN-OMP [50]. The noiseless case provides a fair comparison between the recovery performance of the S-NNLS solvers since, in this scenario, algorithm parameters are easy to select. In the noisy setting, tuning of various algorithm parameters leads to differing behavior of the algorithms, making it difficult to draw wide-reaching conclusions about the relative performance of the algorithms. Therefore, noisy simulations are meant for robustness analysis of R-SBL with DA rather than comparison with other methods. Note that we use terms \mathcal{N}^R and RG interchangeably to denote the rectified Gaussian distribution.

A. Experiment Design

For the MCMC simulations we focus on two problems with different sizes: I.

- P1: $\mathbf{x}^{gen} \in \mathbb{R}_+^{20}$ and $\Phi \in \mathbb{R}^{20 \times 80}$
- P2: $\mathbf{x}^{gen} \in \mathbb{R}_+^{40}$ and $\Phi \in \mathbb{R}^{40 \times 160}$

with elements of \mathbf{x}^{gen} drawn from the RG (Location: 0, Scale: 1) density. For P1 we perform 250 trials for a fixed cardinality of $K = 10$ and for P2 we perform 80 trials for a fixed cardinality of $K = 20$.

To test the proposed DA approach, we generate sparse ground truth solutions $\mathbf{x}^{gen} \in \mathbb{R}_+^{100}$ such that $\|\mathbf{x}^{gen}\|_0 = K$. We draw elements of \mathbf{x}^{gen} according to the following distributions:

- 1) RG (Location: 0, Scale: 1)
- 2) NN-Cauchy (Location: 0, Scale: 1)
- 3) NN-Laplace (Location: 0, Scale: 1)
- 4) NN-Gamma (Location: 1, Scale: 2)
- 5) Chi-square with $\nu = 2$
- 6) Bernoulli with $p(0.5) = 1/2$ and $p(1.5) = 1/2$

where the prefix 'NN' stands for non-negative. The non-negative distributions are obtained by taking the absolute value of the respective probability densities, i.e. $\text{NN-}\mathbf{X} = |\mathbf{X}|$. Next, we randomly generate $\Phi \in \mathbb{R}^{100 \times 400}$ according to the following densities

- I. Gaussian (Location: 0, Scale: 1)
- II. ± 1 with $p(1) = 1/2$ and $p(-1) = 1/2$

and we normalize the columns of Φ to have unit ℓ_2 norm. For a given Φ and \mathbf{x}^{gen} , we compute the synthetic measurements $\mathbf{y} = \Phi \mathbf{x}^{gen}$ and use various S-NNLS algorithms to approximate \mathbf{x}^{gen} by $\hat{\mathbf{x}}$. For the DA simulations, we perform 1000 trials for each combination of distribution for \mathbf{x}^{gen} , distribution for Φ , and cardinality $K = \{1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$.

B. Performance Metrics

To evaluate the performance of S-NNLS algorithms, we use the mean squared recovery error (MSE) and the probability of error in the recovered support (PE), which are two commonly used performance metrics in the compressive sensing literature

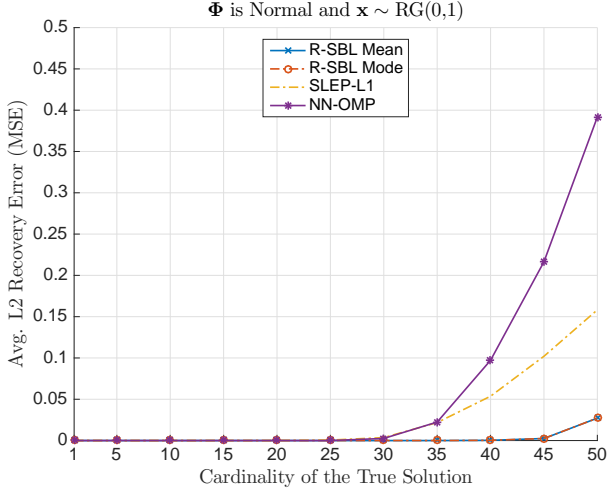


Fig. 2: Average MSE versus cardinality. Φ is $\mathcal{N}(0, 1)$ and x^{gen} is $\text{RG}(0, 1)$

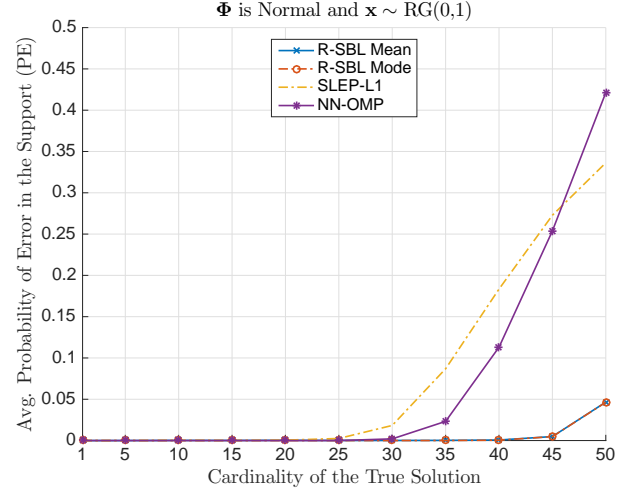


Fig. 3: Average PE versus cardinality. Φ is $\mathcal{N}(0, 1)$ and x^{gen} is $\text{RG}(0, 1)$

[15]. We measure the MSE between the recovered signal \hat{x} and the ground truth x^{gen} using

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i^{gen})^2 \quad (28)$$

Denoting the support of the true solution as S and the support of \hat{x} as \hat{S} , PE is calculated using

$$\text{PE} = \frac{\max\{|S|, |\hat{S}|\} - |S \cap \hat{S}|}{\max\{|S|, |\hat{S}|\}}. \quad (29)$$

A value of $\text{PE} = 0$ indicates that the ground truth and recovered supports are the same, whereas $\text{PE} = 1$ indicates no overlap between supports. Averaging the PE over multiple trials gives us the empirical probability of making errors in the support.

We calculate the average values of MSE and PE over the respective trials, i.e. 250 (P1) and 80 (P2) and 1000 (DA). In the following sections, we use MSE and PE to indicate the averaged values.

C. Experiment Results

In this section we detail the experimental results, showing the benefits of R-SBL compared to other methods. The results will show that, in all of the sparse recovery experiments described, the proposed R-SBL method performs significantly better than its SLEP- ℓ_1 and NN-OMP counterparts both in terms of MSE and PE both in DA and MCMC simulations.

	P1		P2	
	MSE	PE	MSE	PE
Mean R-SBL (MCMC)	0.0692	0.1148	0.0683	0.0853
Mode R-SBL (MCMC)	0.0692	0.0936	0.0687	0.0847
Mean R-SBL (DA)	0.1045	0.1192	0.0885	0.0927
Mode R-SBL (DA)	0.1047	0.0956	0.0886	0.0873
SLEP	0.2013	0.3276	0.1561	0.2947
NN-OMP	0.4824	0.4016	0.3763	0.3380

TABLE I: MCMC and DA Results

Markov-Chain Monte-Carlo Results: Table I shows the MCMC simulation results together with the DA and other baseline methods for problems P1 and P2. As expected R-SBL with both MCMC and DA achieves greater performance with respect to NN-OMP and SLEP, proving the superiority of R-SBL over other S-NNLS solvers. Moreover, since MCMC solution is exact, it achieves better performance compared to the DA. Despite MCMC EM is theoretically more pleasing, numerically and complexity-wise it is not feasible especially for large problem sizes. Trials for P1 took about couple of minutes to hours depending on how ill-posed the problem was, we experienced that simulations for the larger problem P2 took significantly longer. Therefore, we refer the DA as a simple and complexity-wise better alternative to the MCMC method.

Diagonal Approximation (DA) Results: Figures 2 and 3 show the MSE and PE as a function of solution cardinality using the DA approach, with elements of Φ drawn from a normal distribution, $\mathcal{N}(0, 1)$, and x_i^{gen} drawn from $\text{RG}(0, 1)$, and mean and mode refer to the type of point estimate used. We see that R-SBL performs significantly better than other methods. In fact, the average reconstruction error is less than 1% with R-SBL for cardinalities less than $K = 45$. Compared to the other algorithms, this is a significant improvement in reconstruction performance. Also note that, mean and mode point estimates have the same recovery performance.

We find that our R-SBL with DA performs the best regardless of the distribution of x^{gen} and outperforms other algorithms with a large margin. We summarize the MSE and PE for all other distributions of x^{gen} in Table II for $\Phi \sim \mathcal{N}(0, 1)$ and Table III for $\Phi \sim \pm 1$ for a fixed cardinality of $K = 50$. We select $K = 50$ because it is the most difficult experimental setting and because the relative performance of the tested algorithms is largely the same across different experimental settings as compared to Figures 2 and 3.

Table II and Table III show that the R-SBL with DA has superior recovery performance. The MSE for NN-Cauchy, NN-Laplace, Gamma and Chi-square are below 0.6% with

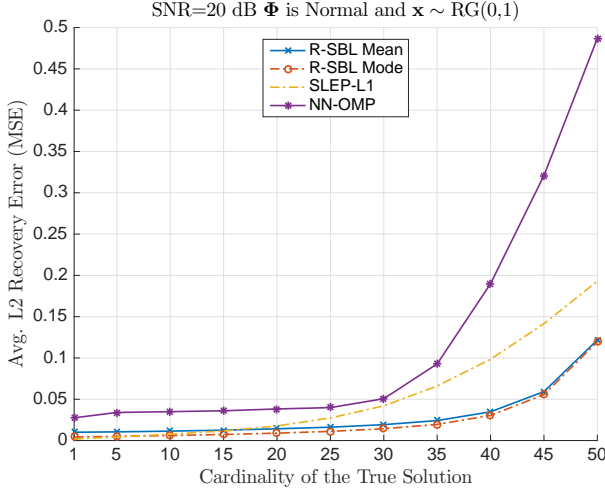


Fig. 4: Average MSE versus cardinality. Φ is $\mathcal{N}(0, 1)$ and \mathbf{x}^{gen} is $\text{RG}(0, 1)$ and \mathbf{v} is Gaussian. SNR is 20 dB.

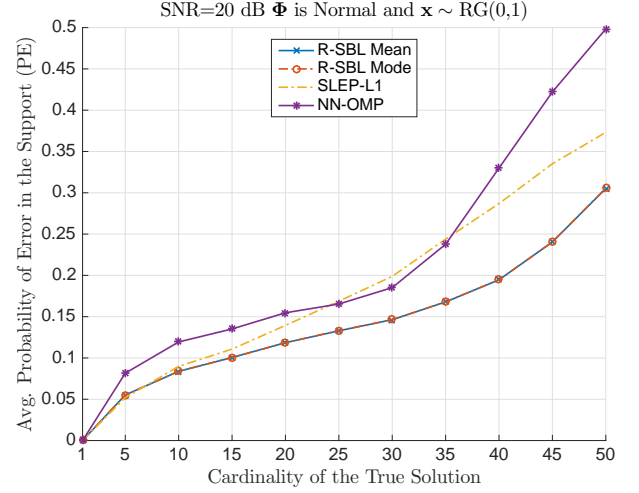


Fig. 5: Average PE versus cardinality. Φ is $\mathcal{N}(0, 1)$ and \mathbf{x}^{gen} is $\text{RG}(0, 1)$ and \mathbf{v} is Gaussian. SNR is 20 dB.

the proposed method, whereas with SLEP and NN-OMP the MSE is above 7%. Similar performance gains are observed in the PE. We also observe that the distribution of Φ does not affect the results significantly and R-SBL with DA is robust for different dictionary distributions.

The worst performance is observed when \mathbf{x}^{gen} is Bernoulli with $p(x_i^{gen} = 0.5) = 1/2$ and $p(x_i^{gen} = 1.5) = 1/2$. We see significantly larger MSE values with the greedy NN-OMP method, while SLEP- ℓ_1 performs significantly better than NN-OMP. R-SBL performs much better compared to the other techniques, although the MSE and PE is relatively high compared to the results for the other distributions of \mathbf{x}^{gen} . This performance loss may result from the fact that the Bernoulli distribution is not approximated well by continuous sparsity-enforcing distributions.

Noise Performance: We now demonstrate the performance of the proposed DA approach in the presence of Gaussian noise, due to computation complexity noise analysis for MCMC is not pursued further. The zero mean noise vector \mathbf{v} is added to measurements \mathbf{y} such that SNR is 20 dB. Selection of tuning parameters for NN-OMP, SLEP- ℓ_1 , and R-SBL is not straightforward in the noisy simulations. To find a common ground between selecting these tuning parameters, we do line a search over the parameters for SLEP and NN-OMP and select

the tuning-parameters that give the smallest MSE error over all cardinalities. For the R-SBL, we fix σ^2 to be the true noise variance. It should be noted that R-SBL is sensitive to the choice of σ^2 and estimation of σ^2 is a topic of research in its own. Depending on the application, σ^2 can be estimated offline or estimation of σ^2 can be incorporated into the R-SBL framework [31].

Figures 4 and 5 show the MSE and PE for Mean R-SBL and Mode R-SBL. \mathbf{x}_i^{gen} is drawn from $\text{RG}(0, 1)$ and the elements of Φ are drawn from $\mathcal{N}(0, 1)$. Compared to Figures 2 and 3, we see that noise degrades the recovery performance of each method. However, both Mean R-SBL and Mode R-SBL perform much better compared to SLEP- ℓ_1 and NN-OMP, especially at higher cardinalities.

Timing Performance: In this part we comment on the timing performance of the R-SBL with DA approach. The R-SBL with DA is computationally much efficient compared to MCMC, which requires drawing samples and simulating the posterior density in an iterative fashion. In fact, complexity-wise R-SBL with DA updates in Figure 1 differs from the SBL introduced in [31] only in first and moment calculations presented in Equations (26) and (24). If further computational efficiency is desired one way of achieving performance speed up is by pruning γ when its elements become close to zero.

	Distribution of \mathbf{x}^{gen}	NN-OMP	SLEP	R-SBL
Avg. MSE	RG(0, 1)	0.3918	0.1582	0.0271
	NN-Cauchy(0, 1)	0.0093	0.0080	0.0002
	NN-Laplace(0, 1)	0.1163	0.0768	0.0055
	Gamma(1, 2)	0.1293	0.0783	0.0047
	Chi-square ($\nu = 2$)	0.1261	0.0746	0.0029
	Bernoulli	0.8972	0.2847	0.2108
Avg. PE	RG(0, 1)	0.4213	0.3363	0.0465
	NN-Cauchy(0, 1)	0.1932	0.3121	0.0394
	NN-Laplace(0, 1)	0.2579	0.3197	0.0169
	Gamma(1, 2)	0.2770	0.3191	0.0147
	Chi-square ($\nu = 2$)	0.2672	0.3210	0.0117
	Bernoulli	0.5473	0.3504	0.2631

TABLE II: Averaged MSE and PE Performance over 1000 trials for cardinality of $K = 50$. Φ is distributed as $\mathcal{N}(0, 1)$.

	Distribution of \mathbf{x}^{gen}	NN-OMP	SLEP	R-SBL
Avg. MSE	RG(0, 1)	0.4028	0.1581	0.0028
	NN-Cauchy(0, 1)	0.0097	0.0083	0.0001
	NN-Laplace(0, 1)	0.1270	0.0794	0.0053
	Gamma(1, 2)	0.1219	0.0740	0.0035
	Chi-square ($\nu = 2$)	0.1318	0.0759	0.0048
	Bernoulli	0.9083	0.2847	0.2086
Avg. PE	RG(0, 1)	0.4202	0.3362	0.0510
	NN-Cauchy(0, 1)	0.1936	0.3144	0.0337
	NN-Laplace(0, 1)	0.2756	0.3256	0.0163
	Gamma(1, 2)	0.2717	0.3195	0.0113
	Chi-square ($\nu = 2$)	0.2658	0.3179	0.0128
	Bernoulli	0.5540	0.3488	0.2624

TABLE III: Averaged MSE and PE Performance over 1000 trials for cardinality of $K = 50$. Φ is distributed as ± 1 .

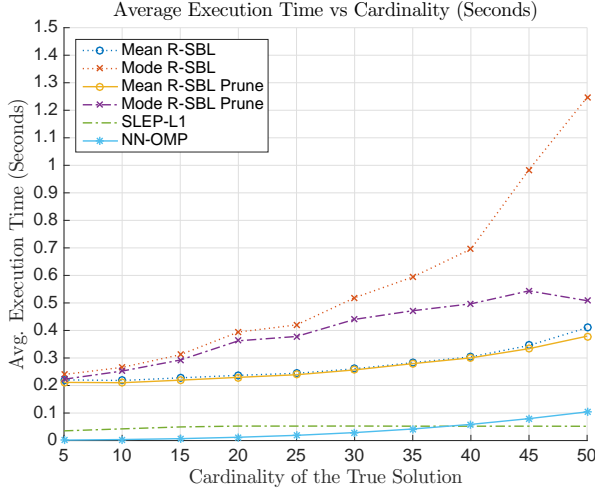


Fig. 6: Average Execution Time to Solve the S-NNLS Problem

When an element γ_i becomes smaller than a threshold i.e. $\epsilon_\gamma = 10^{-5}$, we shrink γ and Φ by ignoring the index i in the next iterations. This effectively reduces the problem dimensions and improves execution time. Alternatively, we can relax the convergence condition by making the ϵ in Algorithm in Figure 1 larger, depending on the desired recovery MSE and PE.

We note that SLEP $_{\ell_1}$ and NN-OMP will still run faster than R-SBL, as OMP and ℓ_1 minimization approaches are faster than SBL, despite the speed-up considerations. Nevertheless, it should be noted that message passing techniques provide considerable speed-up in compressed sensing problems [51], [52], [53] and recent developments in message passing techniques for SBL [54] could be extended to the R-SBL framework, leading to significant speed-up.

Figure 6 shows an example of average execution times to solve the S-NNLS problem. This example includes both the pruning and the no-pruning methods. In the pruning method, we prune γ with a threshold of $\epsilon_\gamma = 10^{-3}$ and the convergence condition is set to $\epsilon = 10^{-5}$ in both cases. Results show that R-SBL with DA is slightly slower than SLEP $_{\ell_1}$ and NN-OMP about 200 milliseconds in average. Pruning helps a lot with the mode point estimate a lot but is rather ineffective for the mean point estimate.

V. CONCLUSION

In this paper, we introduced a hierarchical Bayesian method to solve the S-NNLS problem. We developed the rectified Gaussian scale mixture model as a general and versatile prior to promote sparsity. We constructed the R-SBL algorithm using the EM framework using both the MCMC approach and an approximation technique that is fast, scalable for large problems and effective in performance. We presented two point estimate methods for the EM framework. We demonstrated that R-SBL outperforms the available S-NNLS solvers by a large margin both in terms of signal and support recovery. The performance gains achieved by R-SBL with Diagonal Approximation are consistent across different non-negative data distributions for \mathbf{x} and different sensing matrix

distributions for Φ . Under noisy conditions, we have shown that R-SBL with DA is very robust in the MSE sense with both mean and mode point estimates and achieves a much better reconstruction performance compared to the other methods tested.

REFERENCES

- [1] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1974, vol. 161.
- [2] B. M. Jedyun and S. Khudanpur, "Maximum likelihood set for estimating a probability mass function," *Neural computation*, vol. 17, no. 7, pp. 1508–1530, 2005.
- [3] J. P. Vila and P. Schniter, "An empirical-bayes approach to recovering linearly constrained non-negative sparse signals," *Signal Processing, IEEE Transactions on*, vol. 62, no. 18, pp. 4689–4703, 2014.
- [4] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with 0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, 2012.
- [5] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.
- [6] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, 2008.
- [7] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, "Text mining using non-negative matrix factorizations," in *SDM*, vol. 4, 2004, pp. 452–456.
- [8] V. Monga and M. K. Mihçak, "Robust and secure image hashing via non-negative matrix factorizations," *Information Forensics and Security, IEEE Transactions on*, vol. 2, no. 3, pp. 376–390, 2007.
- [9] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005.
- [10] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.
- [11] P. Sajda, S. Du, T. R. Brown, R. Stoyanova, D. C. Shungu, X. Mao, and L. C. Parra, "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 12, pp. 1453–1465, 2004.
- [12] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, "Sparsity and compressed sensing in radar imaging," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1006–1020, 2010.
- [13] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *INTERSPEECH*, 2012.
- [14] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly, "kt sparse: High frame rate dynamic mri exploiting spatio-temporal sparsity," in *Proceedings of the 13th Annual Meeting of ISMRM, Seattle*, vol. 2420, 2006.
- [15] M. Elad, *Sparse and Redundant Representations*. Springer New York, 2010.
- [16] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge University Press, 2012.
- [17] Y. C. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.
- [18] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.
- [19] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.
- [20] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [21] C.-b. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [22] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

- [23] P. D. Grady and S. T. Rickard, "Compressive sampling of non-negative signals," in *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*. IEEE, 2008, pp. 133–138.
- [24] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*. IEEE, 2008, pp. 3869–3872.
- [25] R. Giri and B. D. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures," *arXiv preprint arXiv:1507.05087*, 2015.
- [26] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.
- [27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Iteratively reweighted least squares for linear regression when errors are normal/independent distributed," 1980.
- [29] K. Lange and J. S. Sinsheimer, "Normal/independent distributions and their applications in robust regression," *Journal of Computational and Graphical Statistics*, vol. 2, no. 2, pp. 175–198, 1993.
- [30] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.
- [31] D. P. Wipf and B. D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704–3716, 2007.
- [32] J. A. Palmer, "Variational and scale mixture representations of non-gaussian densities for estimation in the bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation," 2006.
- [33] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf, "Variational em algorithms for non-gaussian latent variable models," in *Advances in neural information processing systems*, 2005, pp. 1059–1066.
- [34] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998, vol. 31.
- [35] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable bayesian models for promoting sparsity," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6236–6255, 2011.
- [36] R. Schachtner, G. Poeppel, A. Tomé, and E. Lang, "A bayesian approach to the lee-seung update rules for nmf," *Pattern Recognition Letters*, vol. 45, pp. 251–256, 2014.
- [37] M. Harva and A. Kabán, "Variational learning for rectified factor analysis," *Signal Processing*, vol. 87, no. 3, pp. 509–527, 2007.
- [38] J. W. Miskin, "Ensemble learning for independent component analysis," in *Advances in Independent Component Analysis*. Citeseer, 2000.
- [39] M. A. Figueiredo, "Adaptive sparseness for supervised learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1150–1159, 2003.
- [40] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [41] Z. Zhang, T.-P. Jung, S. Makeig, Z. Pi, and B. Rao, "Spatiotemporal sparse bayesian learning with applications to compressed sensing of multichannel physiological signals," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 22, no. 6, pp. 1186–1197, 2014.
- [42] W. C. Horrace, "Some results on the multivariate truncated normal distribution," *Journal of Multivariate Analysis*, vol. 94, no. 1, pp. 209–221, 2005.
- [43] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.
- [44] Y. Li and S. K. Ghosh, "Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints," *Journal of Statistical Theory and Practice*, vol. 9, no. 4, pp. 712–732, 2015.
- [45] M. Magdon-Ismail and J. T. Purnell, "Approximating the covariance matrix of gmms with low-rank perturbations," in *Intelligent Data Engineering and Automated Learning—IDEAL 2010*. Springer, 2010, pp. 300–307.
- [46] W. C. Horrace, "On ranking and selection from independent truncated normal distributions," *Journal of Econometrics*, vol. 126, no. 2, pp. 335–354, 2005.
- [47] J. W. Miskin, "Ensemble learning for independent component analysis," in *Advances in Independent Component Analysis*. Citeseer, 2000.
- [48] B. Hajek, *Random Processes for Engineers*. Cambridge University Press, 2015.
- [49] J. Liu, S. Ji, J. Ye *et al.*, "Slep: Sparse learning with efficient projections," *Arizona State University*, vol. 6, p. 491, 2009.
- [50] A. M. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 4813–4820, 2008.
- [51] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *Signal Processing, IEEE Transactions on*, vol. 61, no. 19, pp. 4658–4672, 2013.
- [52] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.
- [53] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Information Theory Workshop (ITW), 2010 IEEE*. IEEE, 2010, pp. 1–5.
- [54] M. Al-Shoukairi and B. Rao, "Sparse bayesian learning using approximate message passing."